

TOWARD AI-SUPPORTED DECISION MAKING IN INNOVATION: REQUIREMENTS AND DESIGN IMPLICATIONS FOR LLM-BASED MULTI-AGENT SYSTEMS

Bhupender Kumar Saini*

Fraunhofer Institute for Industrial Engineering, Stuttgart
bhupender-kumar.saini@iao.fraunhofer.de

Leonie Isabel Gürtler

Fraunhofer Institute for Industrial Engineering, Stuttgart
leonie.guertler@iao.fraunhofer.de

Daniel Auspurg

Fraunhofer Institute for Industrial Engineering, Stuttgart
daniel.auspurg@iao.fraunhofer.de

Chandan Kumar

Fraunhofer Institute for Industrial Engineering, Stuttgart
chandan.kumar@iao.fraunhofer.de

Matthias Peissner

Fraunhofer Institute for Industrial Engineering, Stuttgart
matthias.peissner@iao.fraunhofer.de

Katharina Hölzle

Fraunhofer Institute for Industrial Engineering, Stuttgart
katharina.hoelzle@iao.fraunhofer.de

Abstract

Innovation decisions in product development are frequently made under conditions of incomplete information, fragmented documentation, and limited cross-functional coordination. These are also the conditions under which overreliance on AI outputs is most likely, particularly when the system's underlying uncertainty is not clearly communicated. This study examines what decision makers in innovation contexts require from AI-based decision support and what this implies for the design of large language model (LLM) based multi-agent systems. An on-line survey ($n = 33$) of participants from twelve German organizations was combined with 16 semi-structured interviews across fourteen firms. The findings indicate that practitioners face recurring breakdowns in information integration, reasoning traceability, and cross-functional alignment, while consistently demanding transparency, human oversight, explicit handling of data gaps, and secure treatment of sensitive information. Based on these findings, seven design requirements and six design principles for AI-supported decision systems in innovation contexts are derived and their implications for LLM-based multi-agent architectures are discussed. The central contribution is an empirically grounded framework that positions AI as an analytical partner rather than an autonomous decision maker, identifies when multi-agent configurations are appropriate for coordinating distinct functional perspectives, and proposes that such systems are most likely to be effective when uncertainty is made explicit, adoption is

¹Corresponding author

introduced incrementally, and AI capabilities are treated as complementary to human expertise rather than as a substitute for it.

Keywords: AI-supported decision making; LLM-based multi-agent systems; innovation management; product development; design requirements; human-AI collaboration

1. INTRODUCTION

Strategic decisions in innovation and product development are among the most consequential choices organizations face. They determine which products reach the market, how resources are allocated across projects, and whether a firm can sustain its competitive position over time (Schuh et al., 2011). Yet these decisions are rarely made under stable or fully knowable conditions. Relevant information is often incomplete, distributed across functions, weakly documented, or held only as tacit experience by individual experts. In practice, innovation decisions frequently combine formal analysis with intuition, negotiation, and time-constrained judgment.

Traditional decision support tools such as SWOT analysis, multi-criteria decision methods, and cost-benefit frameworks were designed for well-defined problems with stable criteria and bounded alternatives. Innovation decisions rarely fit this description: criteria evolve as markets shift, evaluation parameters change as new information surfaces, and the boundaries of the problem itself are often unclear (Varnes, 2025). Practitioners, therefore, frequently rely on experience and informal consultation rather than systematic analysis.

Recent advances in generative AI have renewed interest in AI-supported decision making, particularly in contexts where relevant knowledge is embedded in unstructured text, distributed documents, and heterogeneous stakeholder inputs. Large language model (LLM) based multi-agent systems are especially attractive in this setting because they can distribute analytical tasks across specialized roles and make multi-perspective reasoning more explicit. However, the practical promise of such systems in innovation management remains under-specified. It is still unclear which practitioner needs are specific enough to justify a multi-agent architecture, which requirements apply to AI-supported decision systems more broadly, and how such systems should be designed so that human judgment remains meaningfully in control.

Current AI decision-support systems are often difficult to trust in practice. They may produce recommendations without making their reasoning visible, obscure the limits of the underlying evidence, or reduce user oversight when automation is presented with unwarranted confidence. Trust, as used in this study, refers to reliance that is appropriately calibrated to the system's actual reliability, not uncritical compliance with AI outputs. Research on AI-assisted decision making has repeatedly shown that transparency, explainability, and user control are important conditions for effective human-AI collaboration, and that many existing systems do not adequately meet these requirements (Steyvers and Kumar, 2024; Romeo and Conti, 2025). Empirically grounded design knowledge linking the needs of innovation practitioners to the design of LLM-based multi-agent systems remains limited.

The present study addresses this gap by investigating the following research questions:

- **RQ1:** What are the key challenges and requirements that decision makers face in strategic innovation decisions?
- **RQ2:** What design requirements should guide AI-supported decision systems for innovation decision support, and which of these have particular implications for trustworthy LLM-based multi-agent systems?
- **RQ3:** What design principles can help keep such systems aligned with human judgment and organizational practices?

To address these questions, a two-phase mixed-method study was conducted: an online

survey ($n = 33$) and semi-structured interviews ($n = 16$) with German practitioners (Section 3). Following a review of the theoretical background (Section 2), empirical findings are presented (Section 4), from which seven design requirements (DR1–DR7, Section 5) and six design principles (DP1–DP6, Section 6) are derived. The paper concludes with a discussion of implications for LLM-based multi-agent architectures, study limitations, and future research directions (Sections 7 and 8). The contribution is empirically derived design guidance, not a technically implemented system.

2. THEORETICAL BACKGROUND

2.1 Decision Making in Innovation

Innovation decisions differ from routine operational choices in important ways. They unfold in the early, uncertain phases of development, where ambiguity is high, criteria evolve, and information about technologies, markets, and customer needs is incomplete (Schuh et al., 2011). Herbert Simon’s concept of bounded rationality (Simon, 1955) offers a useful starting point: rather than optimizing, decision makers *satisfice*, selecting options that meet a minimum acceptable threshold given constraints of time, information, and cognitive capacity. Schwarz et al. (2022) extended this idea to AI, noting that while AI can process information at scale, AI systems are themselves bounded by the data they access, a point with direct implications for how AI-based decision support should be designed.

Beyond individual cognition, innovation decisions typically involve multiple stakeholders with competing priorities. Outcomes may be influenced more by seniority than by evidence, a pattern known as the HiPPO effect (Highest Paid Person’s Opinion) (Varnes, 2025). Reasoning chains are rarely documented, which impedes organizational learning and means that each new decision is effectively made from scratch. Together, these challenges, cognitive limits, stakeholder conflict, and absent decision memory, define the problem space that AI-based decision support needs to address. Growing recognition of these challenges has led to interest in AI as a structural component of innovation management (Füller et al., 2024; Bouschery et al., 2023), moving the conversation from AI as an efficiency tool to AI as a partner in strategic decision making.

2.2 Decision Support Systems

Traditional decision support tools such as SWOT analysis, multi-criteria decision making, and scenario planning work well when problems are clearly defined and criteria are stable (Witkowski and Wodecki, 2025). Innovation decisions routinely violate these assumptions, which helps explain why practitioners in the present study described existing tools as too rigid and insufficiently adaptive. AI-augmented decision making (AIADM) offers new possibilities by enabling systems to adapt to context and actively participate in evaluation rather than simply presenting static outputs (Herath Pathirannehelage et al., 2025). When AI handles information synthesis tasks, human experts can focus on creative and evaluative work, and larger problem and solution spaces can be explored (Bouschery et al., 2023).

In entrepreneurial and early-stage innovation contexts, generative AI appears to support rather than replace human judgment, provided the interaction is structured to preserve human agency (Bastone et al., 2026). Generative AI has also been found to narrow the performance gap between novices and experienced practitioners in structured planning tasks (Hettrich et al., 2025), suggesting potential for more equitable decision support across organizational levels. Despite this potential, a substantial gap persists between experimentation and embedded use: Saup et al. (2026) describes a “pilot-to-decision gap” in which organizations try generative AI but fail to integrate it into core decision workflows, largely due to governance challenges and

insufficient trust. Closing this gap requires not just better technology but design principles that keep humans meaningfully in control (Mariani and Dwivedi, 2024).

2.3 Large Language Models and Multi-Agent Systems

LLMs can process unstructured text, engage in multi-step reasoning, and maintain context across interactions (Chiarello et al., 2024), which makes them well suited to innovation decision support where relevant knowledge is embedded in documents and reports. Three system levels are relevant to this study (Guo et al., 2024a; Hosseini and Seilani, 2025): an *AI agent* that perceives its environment and acts autonomously; a *multi-agent system* (MAS) in which agents coordinate on tasks no single agent could handle; and an *LLM-based multi-agent system* (LLM-MAS) in which agents are powered by language models, enabling flexible reasoning and conversational interaction. The LLM-MAS design space is maturing rapidly, and applied work suggests that multi-agent workflows can help structure complex analytical tasks in ways that are directly relevant to decision support (Guo et al., 2024a; Hosseini and Seilani, 2025; Sun et al., 2025; Krueger et al., 2025; Sami et al., 2026).

Assigning agents to distinct functional roles mirrors the multi-stakeholder complexity described in Section 2.1. Role specialization can help structure complex deliberation, but it also introduces coordination and evaluation challenges that must be managed deliberately (Sun et al., 2025). Krueger et al. (2025) shows that LLM-based systems can help translate practitioner knowledge into structured product requirements, a challenge closely related to the information-structuring demands identified in this study. Sami et al. (2026) find that structuring the human-LLM interface in multi-agent requirements analysis improves both output quality and user confidence. Across this body of work, aggregation, critique, and human-approval checkpoints emerge as important mechanisms for reliability and control in multi-agent settings (Guo et al., 2024a; Sun et al., 2025).

2.4 Responsible and Transparent AI

Trust in AI decision support rests on transparency, credibility, and alignment with user expectations. Users need not just a recommendation but the reasoning behind it: the evidence considered, the assumptions made, and the trade-offs involved. The performance benefits of intelligent decision support depend on a match between the user’s skill level and the system’s design, which underscores the importance of adaptive autonomy (Peng et al., 2024). A key risk is automation bias, the tendency to over-rely on automated outputs. Low AI literacy and poorly designed explanations increase uncritical compliance (Romeo and Conti, 2025), and automation bias has been observed even among experienced users in decision support contexts (Kücking et al., 2024). Guo et al. (2024b) shows formally that appropriate AI reliance can be defined as a function of task uncertainty and AI reliability, with direct implications for how confidence indicators should be communicated. Steyvers and Kumar (2024) identifies three core challenges for effective human-AI decision making: knowing when to trust the AI, understanding its limits, and maintaining meaningful oversight.

As AI systems become more agentic, the risk profile expands. Chan et al. (2023) identify harms specific to agentic systems, including goal misalignment and compounding errors across multi-step decision chains, risks that are particularly relevant in LLM-MAS architectures. Yang and Yang (2024) documents broader social risks, including the erosion of human accountability, when AI is positioned as an active participant in organizational decisions.

A central tension runs through this literature. LLM-based multi-agent systems are designed to help practitioners work with incomplete information, yet they can generate confident-sounding outputs even when the underlying evidence is weak. This study refers to this as

the *incomplete information tension*: AI support is most needed when data are scarce, yet a system that conceals its own gaps may mislead users rather than assist them. This tension, which echoes Schwarz et al. (2022)’s point that AI operates under bounded intelligence, directly shapes the design requirements developed in this study.

2.5 Research Gap and Positioning

This study sits at the intersection of innovation management, AI-augmented decision making, and multi-agent systems design. It takes a design-oriented approach: empirical findings from practitioners are used to derive system requirements and principles. This distinguishes it from largely technical LLM-MAS studies that optimize for performance without grounding in user needs (Guo et al., 2024a), and from high-level AI governance reviews that do not address concrete multi-agent design (Mariani and Dwivedi, 2024). A gap remains in empirically grounded design knowledge that links what innovation practitioners need to how an LLM-MAS should be built. This study addresses that gap by keeping practitioner experience at the center of the research process. It extends the AIADM design principles developed by Herath Pathirannehelage et al. (2025) through action design research by applying and specifying them for multi-agent architectures in innovation management contexts.

3. METHODOLOGY

This study uses a two-phase mixed-method design in which quantitative data collection precedes and informs qualitative inquiry (Creswell and Creswell, 2018). This approach was chosen because the research question calls for both broad coverage across many practitioners and a deeper understanding of the reasoning and context behind their experiences, which neither method alone is likely to provide adequately (Johnson and Onwuegbuzie, 2004). The two phases are shown in Figure 1. In Phase 1, an online survey was used to map the frequency and distribution of decision-making challenges, current tool use, and expectations about AI across a diverse cross-industry sample. This phase was designed as an exploratory step rather than a formal statistical test, serving to identify which themes appeared most widely before the qualitative phase began. In Phase 2, semi-structured interviews were conducted to explain the mechanisms, contextual factors, and practitioner reasoning behind the patterns observed in the survey. A final synthesis step mapped the combined findings onto system design requirements. This sequential structure means that the qualitative inquiry was guided by the quantitative results rather than being conducted independently of them.

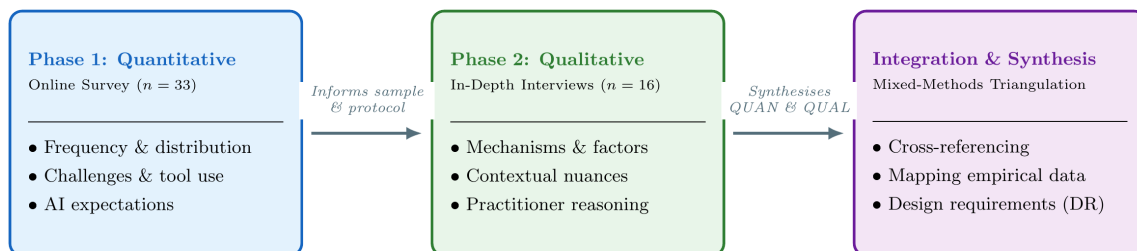


Figure 1: Overview of the two-phase mixed-method research design

Both studies target professionals who bear responsibility for strategic or operational decisions in German organizations engaged in innovation and product development. This population is appropriate because such practitioners are the intended primary users of the LLM-MAS concept investigated here.

3.1 Study 1: Online Survey

The survey was conducted within the *InnoHub Generative AI*, a structured cross-company innovation collaboration organized by Fraunhofer IAO and the University of Stuttgart IAT. Participants were recruited through the consortium network of 12 partner organizations spanning manufacturing, technology, consulting, financial services, and healthcare sectors, ensuring deliberate coverage of diverse organizational contexts. Out of the 46 invited participants, 33 provided valid responses, corresponding to a response rate of 71.7 %. Respondents held roles including HR managers, AI coordinators, digital transformation specialists, product managers, and heads of quality, sales, and innovation, ensuring broad functional coverage. Table 1 summarizes their background characteristics.

Table 1: Participant Background Characteristics: Study 1 Online Survey ($n = 33$)

Characteristic	n	%
<i>Self-assessed AI knowledge</i>		
No prior knowledge	9	27
Basic	19	58
Advanced	5	15
<i>Familiarity with AI-based decision support</i>		
Not familiar	24	73
Somewhat familiar	8	24
Very familiar	1	3
<i>Familiarity with generative AI and multi-agent systems</i>		
Not familiar	17	52
Somewhat familiar	14	42
Very familiar	2	6

The survey instrument was developed in three steps. First, the theoretical framework was used to identify the topics to be covered: decision-making frequency and methods, key challenges, current tool use and its limitations, AI interest, knowledge, and familiarity with AI-based systems, concerns about AI adoption, preferred types of AI support, and expectations about future AI use. Second, questions were drafted by the research team and linked to each topic, drawing on established decision-support and technology-acceptance scales where available. Third, the instrument was reviewed by two domain experts external to the research team and piloted with three prospective respondents; their feedback was used to improve the wording and response formats before the survey was distributed. Because the analysis was descriptive in nature, formal reliability statistics such as Cronbach’s α were not computed. To account for differences in technical background among participants, a concise and neutral introduction to generative AI capabilities was included in the questionnaire before questions about future expectations were asked.

The final instrument contained eight thematic blocks. Closed-ended questions used multiple-choice formats, five-point Likert scales, and forced-ranking. Open-ended questions were included to capture views not anticipated by the fixed-choice items, most notably on current tool limitations and future AI expectations. The survey was administered online. Data were analyzed using descriptive statistics and frequency counts; open-ended responses were reviewed thematically to identify patterns not captured by the closed-ended items.

3.2 Study 2: In-Depth Interviews

The second study comprised semi-structured in-depth interviews with 16 decision makers from 14 German firms, selected through purposive sampling to ensure coverage of diverse industries, organizational sizes, and decision types. Table 2 provides an overview of participants.

Table 2: Overview of Interviewed Decision Makers (DMs)

ID	Role / Position	Industry	Specialization
DM1	Team Lead, IT & Digitalization	Technology	Digital Transformation
DM2	Manager, Industrial Design	Product Design	Strategy
DM3	Head of Prototyping & Testing	Automotive	Powertrain
DM4	Innovation Manager	Consumer Goods	Innovation
DM5	CEO, Product Design Agency	Consulting	Design
DM6	Management Consultant	Consulting	Product Development
DM7	Product Manager	Consumer Electronics	Product Management
DM8	Chief Technology Officer (CTO)	Medical Devices	Technology Leadership
DM9	Director of User Experience (UX)	Software Platforms	User Experience
DM10	VP, Supply Chain & Operations	Automotive Components	Supply Chain
DM11	Head of Quality Assurance	Pharmaceutical Manufacturing	Quality Assurance
DM12	Senior Product Marketing Manager	B2B SaaS	Product Marketing
DM13	Director, Manufacturing Engineering	Consumer Goods Manufacturing	Manufacturing Engineering
DM14	Head of R&D	Chemical Industry	Research & Development
DM15	Business Development Director	Industrial Equipment	Business Development
DM16	Product Development Lead	Sustainable Packaging	Product Development

Participants were selected through purposive sampling to ensure variation in industry sector, organizational size, decision type, and product development context (Patton, 2015). Each participant was required to hold formal responsibility for at least one type of strategic or operational decision in an innovation or product development setting. Recruitment continued until thematic saturation was considered to have been reached, defined here as the point at which two consecutive interviews did not introduce any substantially new themes or codes. On this basis, no new themes were identified after the twelfth interview. Four additional interviews were then conducted, deliberately selected from different organizational backgrounds, to confirm that the identified themes held across varying contexts.

The interview guide was developed in two stages. In the first stage, the survey findings and the theoretical framework were used to define four main question areas: (1) the participant's role and the types of decisions associated with it; (2) challenges encountered in current decision processes, including the tools used and their limitations; (3) possible use cases for an LLM-based multi-agent system; and (4) requirements and concerns related to AI-based decision support. Follow-up probing questions were prepared for each area, and a summary of the interview protocol is provided in the appendix. In the second stage, the guide was piloted with one participant whose data were then excluded from the main analysis; the reported $n = 16$ and the saturation count therefore refer only to the main interview series. Minor wording adjustments were made following the pilot before the main series began.

Interviews lasted between 45 and 75 minutes, were conducted in German via video call, and were audio-recorded with the participant's consent. Recordings were transcribed in full, and all identifying information was replaced with anonymized identifiers (DM1–DM16) before analysis. All quotations translated into English for this study were reviewed by a bilingual co-author to verify accuracy.

Transcripts were analyzed using thematic analysis following the steps described by Braun

and Clarke (2006): familiarization with the data, initial coding, theme construction, theme review, theme definition, and write-up. Coding was primarily data-driven, although it was guided by the four question areas used in the interview guide. As a quality check, a randomly selected 20 % of the coded transcript segments (approximately 420 coded segments in total, where a segment is defined as a continuous passage assigned a single code) were independently coded by a second researcher. The two coders agreed on 84 % of segments, and the chance-corrected agreement coefficient (Cohen's κ) was 0.78, which is generally considered to indicate substantial agreement (Landis and Koch, 1977). Disagreements were resolved through discussion and by refining theme definitions. The resulting themes were then organized within the Value Proposition Canvas framework (Osterwalder et al., 2014) using the categories of Jobs to be Done, Pains, and Gains, in order to translate practitioner needs into specific system capability requirements.

3.3 Quality, Validity, and Ethics

Several measures were taken to strengthen the reliability and credibility of the study. In the survey, each item was grounded in established literature and reviewed by external experts before distribution. In the qualitative analysis, credibility was supported through a limited member-resonance check: a summary of the identified themes was shared with four interview participants, all of whom indicated that the themes were consistent with their expressed views. The generalizability of the findings is bounded by the sample; the study focuses on German-speaking organizations in innovation-intensive sectors, and the findings should be interpreted within this context.

A further limitation concerns the sampling frame: both studies recruited from or around the *InnoHub Generative AI* consortium, whose members had voluntarily engaged with an AI-focused innovation programme. Participants in such a setting are likely to be more positively oriented toward AI adoption, and possibly more AI-literate, than the wider population of innovation managers. The concern levels and expectations reported here may therefore reflect a more favorable baseline than would be found in a broader or less self-selected sample, and this should be kept in mind when interpreting the findings. The reliability of the qualitative analysis was further supported by the intercoder agreement procedure described above (raw agreement 84 %, Cohen's $\kappa = 0.78$) and by maintaining a full record of all coding decisions.

Both studies were conducted in accordance with applicable data protection regulations. Survey responses were collected anonymously and stored on institution-controlled servers in compliance with the EU General Data Protection Regulation (GDPR). Interview participants received a written information sheet before participation and provided informed consent for recording, transcription, and anonymized use of their statements in research publications. No personally identifiable information appears in this study.

Regarding funding and competing interests, the research reported here was carried out as part of the *InnoHub Generative AI* initiative coordinated by Fraunhofer IAO and the University of Stuttgart IAT. The authors declare that they have no competing financial interests or personal relationships that could be perceived as influencing the work reported in this study.

3.4 Integrating the Two Studies

Patterns identified in the survey were compared with themes from the interviews to assess whether the quantitative results were supported by the qualitative data and vice versa. Themes that appeared strongly in the interviews were checked against the survey to assess how widely they were shared, while survey patterns were examined in the interviews to understand the underlying mechanisms. This process revealed consistent agreement on several key themes,

including the difficulty of managing complex decision criteria, the importance of transparency and trust, and the shared expectation that AI should support rather than replace human judgment.

The combined findings were then organized within the Value Proposition Canvas framework (Osterwalder et al., 2014) to translate practitioner needs into specific system requirements. Each theme was classified as a *Job to be Done* (a task the decision maker needs to accomplish), a *Pain* (a difficulty or barrier in the current process), or a *Gain* (a desired improvement or outcome). For example, the recurring problem of fragmented documentation was classified as a Pain, while the desire for traceable AI reasoning was classified as a Gain linked to the same Job. This structure made the step from empirical observation to design requirement explicit: each requirement was developed to address one or more identified Pains or Gains, rooted in a specific Job. The resulting mapping is summarized in Table 5 in Section 5.

4. FINDINGS

All frequency counts reported in this section refer to the survey sample ($n = 33$) unless stated otherwise, and should be interpreted as characterizing that purposive sample rather than as population estimates.

4.1 Decision-Making Practices and Challenges

Most survey respondents reported that strategic decisions were made monthly, and organizational goals were identified as the primary guiding factor. Regarding methods, reasoning from prior analogous cases¹ was most frequently reported (21 of 33 respondents), followed by data analysis (20), SWOT analysis (16), and the 5-Why method (9). Only one respondent reported using formal multi-criteria methods such as AHP or TOPSIS. Several respondents also mentioned informal approaches, including pro/con lists, intuition, and competitive analysis. Table 3 summarizes these results.

Table 3: Decision-making methods currently used (survey, $n = 33$)

Method	Count
Reasoning from prior analogous cases	21
Data analysis	20
SWOT analysis	16
5-Why analysis	9
Multi-criteria (AHP, TOPSIS, etc.)	1

Note: Multiple selections were allowed, so counts exceed $n = 33$.

Figure 2 shows the distribution of decision-making methods and the frequency of the top five reported challenges. Complex decision criteria were the most commonly cited difficulty (19 of 33), followed by time pressure (18), lack of relevant data (16), inadequate tools (12), and market uncertainty (12).

The interviews reinforced and deepened these findings. DM1 (Team Lead, IT & Digitalization) described recurring poor data quality and a lack of structured evaluation, with experts brought in too late. DM3 (Head of Prototyping & Testing) noted that documentation is frequently fragmented, making it difficult to reconstruct past reasoning chains. DM4 (Innovation Manager) pointed to subjective evidence and senior-level pressure overriding evidence (the HiPPO effect). DM6 (Management Consultant) identified siloed communication across depart-

¹The term is used here in the informal practitioner sense of drawing on experience, not as a reference to the formal case-based reasoning technique.

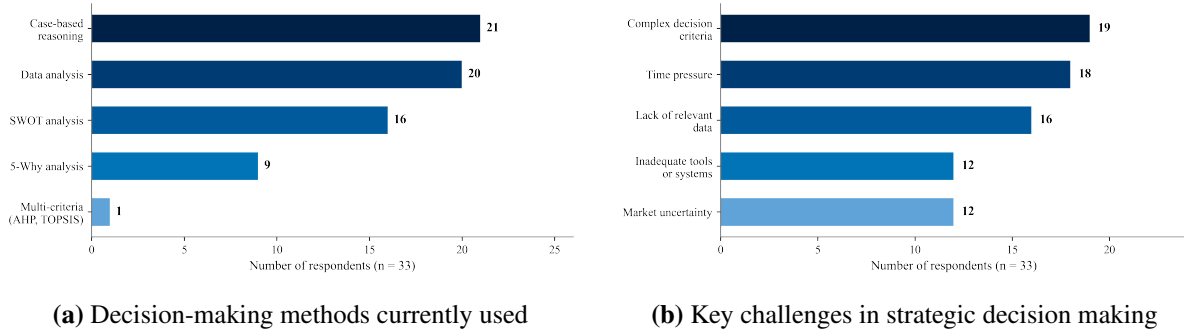


Figure 2: Survey results on current decision-making practice (n = 33): methods currently in use (left) and frequency of reported challenges (right)

ments as a persistent barrier to holistic decision-making. DM7 (Product Manager) reported that feature requests from multiple stakeholders create conflicting priorities that resist systematic resolution.

Together, these accounts point to a consistent pattern: strategic innovation decisions are hampered by information gaps, coordination failures, cognitive biases, and tool limitations that are embedded in organizational structures.

4.2 Tool Landscape and Limitations

Survey respondents reported using tools including Power BI, SAP- and Excel-based solutions, Conceptboard, mind maps, PowerPoint, internal dashboards, and in a few cases, LLMs for structuring data and ideas. The overall picture was one of widespread dissatisfaction: 12 of 33 respondents identified inadequate tools as a key challenge, and open-ended responses emphasized a lack of flexibility and a heavy burden of manual data preparation. Combined with the challenges of insufficient data (16/33) and time pressure (18/33), these findings suggest that the current tool environment amplifies rather than reduces the cognitive demands of innovation decisions.

Interview participants provided specific examples. Data collection was described as largely manual and time-consuming. Analytical tools were seen as rigid, unable to adapt when decision parameters changed. Several interviewees noted that despite having formal tools available, actual decisions frequently relied on experience and intuition, because the effort needed to gather comprehensive information typically exceeded the time available.

4.3 AI Use Cases and Potentials

The interviews revealed a wide range of potential use cases for LLM-MAS. Participants were first asked about their current operational challenges before discussing how AI systems might help, so that the proposed applications reflected genuine practitioner needs. The responses were organized into seven thematic clusters, summarized in Table 4.

A pattern that emerged across many interviews was the potential for LLM-MAS to serve as an “intelligent mediator” (DM6) that systematically collects arguments, translates between departmental technical languages, and helps balance competing perspectives. DM4 envisioned the system as a committee decision support tool capable of reconciling subjective opinions with objective data. DM5 described a liaison role between departments, leveraging access to information from all sides to facilitate balanced decisions.

Three further clusters warrant brief elaboration. In *trend and competitive analysis*, DM7 and DM12 envisioned continuous, automated monitoring of market signals and competitor activity, freeing analysts from manual scanning. In *regulatory and compliance assessment*, DM8 and

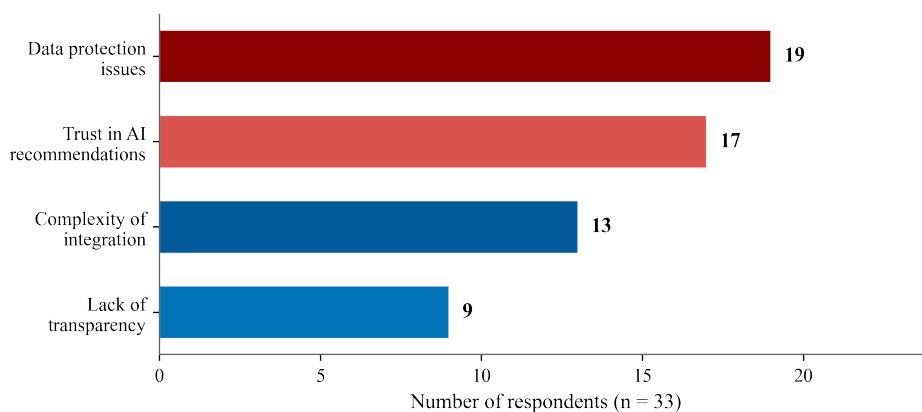
Table 4: Use case categories for LLM-MAS identified from interviews

Use Case Category	Description	DMs
Feasibility & impact assessment	Evaluating viability, scalability, and cost-benefit of new ideas	1, 3
Make-or-buy decisions	Supporting build vs. buy vs. outsource decisions	1, 6, 13
Committee decision support	Reconciling subjective opinions with objective data in group settings	4
Trend & competitive analysis	Continuous monitoring of market trends, competitors, and technologies	4, 7, 12
Cross-functional mediation	Translating between departmental perspectives and aligning stakeholders	5, 6, 9
Regulatory & compliance assessment	Analyzing regulatory requirements, quality risks, and compliance pathways	8, 11
Scenario planning & risk analysis	Modeling disruption scenarios, supply chain risks, and uncertainties	3, 10, 14, 16

DM11 pointed to the potential for an LLM-MAS to track changing regulatory requirements and flag compliance risks in real time, an area where the documentation demands are high and the cost of oversight gaps is severe. In *scenario planning and risk analysis*, DM3, DM10, DM14, and DM16 described the value of a system capable of modelling supply chain disruptions, technology shifts, and demand uncertainty simultaneously, a task that exceeds the capacity of any single analyst. Feasibility and make-or-buy decisions, while mentioned by fewer participants (DM1, DM3, DM6, DM13), were noted as potentially the most tractable entry point because the decision criteria are relatively well-defined and the output is a comparative evaluation rather than an open-ended recommendation.

4.4 Practitioner Requirements and Concerns

Both the survey and the interviews provided detailed insights into what practitioners expect from and worry about in AI-based decision support. Figure 3 presents the survey results on key concerns.

**Figure 3:** Key concerns about AI adoption in decision making (survey, n = 33)

Transparency and traceability was the most consistent theme across both data sources. DM3 stated explicitly: “I want to know where the information the system uses comes from.” DM1, DM4, and DM6 all emphasized that decision logic must be transparent and traceable. In regulated industries, this becomes even more stringent: DM8 (Medical Devices) and DM11 (Phar-

maceutical QA) noted that every decision support interaction must be documented with timestamps and a rationale for audit purposes.

Trust and data protection were the most frequently cited concerns in the survey, with 19 of 33 respondents identifying data protection as a key barrier. DM14 and DM15 both stressed that intellectual property considerations require secure handling of all research and strategic data.

Human oversight was treated as non-negotiable by interview participants across sectors. DM4 emphasized that an LLM-MAS should be introduced as a supportive colleague, not an autonomous decision maker. DM5 argued that AI should assist in decisions but must never become the decision maker itself, and that responsibility must remain with humans.

Gap identification emerged as a distinctive requirement: DM1 insisted that the system must provide clear indications of data gaps or insufficient information quality. DM14 emphasized that the system must acknowledge when insufficient data exists rather than extrapolating unreliably. This gap identification requirement directly addresses the incomplete information tension discussed in Section 2.4.

4.5 Future AI Expectations

Survey respondents expressed the strongest interest in AI-generated insight types that were directly relevant to their decision tasks. When asked to rank the types of insights they would most value, trend analyses ranked highest, followed by risk analysis and competitor analysis. Figure 4 illustrates the full ranking distribution across the five insight types.

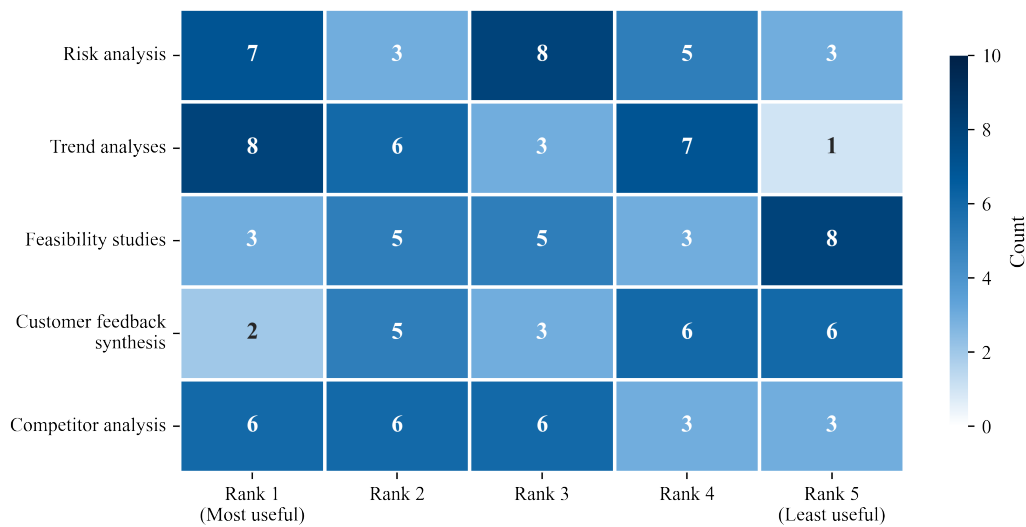


Figure 4: Preferred types of AI-generated insights (survey, n = 33; Rank 1 = most useful, Rank 5 = least useful)

Open-ended responses about the expected role of AI over the next five years revealed a consistent narrative. Participants anticipated that AI would grow into an important supportive tool while humans remain in the lead. Representative statements include: “AI will support strategic decisions; humans decide”; “becoming more important, but humans remain in the lead”; and “AI will grow, but remain a partner for decision consideration.” Others anticipated efficiency gains and cost reductions, while emphasizing that AI should automate repetitive processes rather than replace creative or strategic thinking.

The interviews added further nuance. DM4 emphasized that AI is particularly valuable in the early phases of innovation, such as exploring new subject areas and scanning for emerging technologies, before decisions become locked in. The same participant advocated for itera-

tive introduction, starting with a few clearly defined tasks and expanding scope gradually as trust builds. Several other interviewees stressed that change management is indispensable for successful adoption: technology alone, they argued, is insufficient without a clear process for introducing the system, defining its role, and training people to collaborate with it effectively.

4.6 Synthesis: Three Structural Breakdowns

Taken together, the findings across the preceding subsections point not to isolated pain points but to three structural breakdowns in how innovation decisions are currently made (see also Figure 2b).

The first is *information fragmentation*: data are scattered across departments, rarely integrated, and often never formally documented. The consequence is that every decision draws on an incomplete, unverifiable picture of the relevant facts.

The second is *absent reasoning structure*: without tools that impose discipline on how options are evaluated, decisions default to informal heuristics, seniority-based authority (the HiPPO effect), or accumulated intuition. The evaluation process is rarely traceable ex post.

The third breakdown is the *incomplete information tension* introduced in Section 2.4. Practitioners knew their information was incomplete but still had to act. They wanted AI support that helped them move forward despite gaps, while being concerned that a system confidently filling those gaps with unverifiable content could undermine rather than support their decisions.

These three structural breakdowns are the direct origin of the design requirements developed in Section 5. Each requirement can be traced to one or more of them.

5. DESIGN REQUIREMENTS

The findings in Section 4 reveal a systemic mismatch between the complexity of innovation decisions and the tools practitioners currently have available. Three structural breakdowns recur across the data: information fragmentation, absent reasoning structure, and the need to act under incomplete information without generating false confidence. Each design requirement addresses one or more of these breakdowns. Together, they form a coherent specification for an AI-supported decision system grounded in practitioner experience.

For each requirement, the link between specific practitioner challenges, the corresponding need, and a concrete system capability was traced using the Jobs-to-be-Done, Pains, and Gains structure of the Value Proposition Canvas (Section 3.4). Where a theme appeared independently across multiple interviews, the relevant participant IDs are listed; this mapping was cross-checked against the survey data to confirm that the challenges were broadly shared and not specific to individual respondents. Table 5 summarizes the mapping.

5.1 The Seven Design Requirements

Based on the mapping in Table 5, seven design requirements are proposed for AI-supported decision systems in innovation contexts. Several have particular implications for LLM-based multi-agent configurations, where multiple functional perspectives must be coordinated. Empirical sources are not repeated in full below, as they appear in the table.

DR1: Transparent Reasoning Chains. Every recommendation must be accompanied by a traceable chain of reasoning that shows the evidence considered, the assumptions made, and the logic applied. Sources must be cited, and the reasoning steps must be presented in a way that allows decision makers to follow, verify, and challenge them. This addresses the transparency demands raised in the interviews and the auditability requirements of regulated industries (DM8, DM11), where lack of transparency was identified as a key barrier to AI adoption.

DR2: Multi-Source Data Integration with Gap Identification. The system must be able

Table 5: Mapping empirical findings to design requirements

Empirical Finding	Evidence	Practitioner Need	Req.
Fragmented documentation; lost reasoning chains	DM1, DM3, DM4, DM6	Traceable, verifiable justifications for recommendations	DR1
Poor data quality; incomplete information	DM1, DM10, DM14; Survey (16/33)	Awareness of data gaps and quality issues	DR2
Complex decision criteria; need for structured iterative tracking	DM3; Survey (19/33)	Structured decomposition of multi-criteria problems	DR3
Human oversight non-negotiable	DM4, DM5, DM8, DM11, DM14	User control over AI involvement level	DR4
Conflicting stakeholder interests; HiPPO effect	DM4, DM6, DM7	Structured argumentation and counter-argument capability	DR5
Reliance on intuition; tacit knowledge not captured in data	DM1, DM3, DM4, DM5	Integration of experiential knowledge alongside data	DR6
Data protection as top concern; IP sensitivity	DM14, DM15; Survey (19/33)	Privacy and security as architectural principle	DR7

to integrate information from diverse internal and external sources, including databases, documents, market reports, and technical specifications. Crucially, it must explicitly flag data gaps, quality issues, and cases where the available information is insufficient for a reliable recommendation. This responds to data quality concerns raised by DM1, DM10, and DM14, as well as the challenge of insufficient data reported by 16 of 33 survey respondents.

DR3: Modular and Traceable Decision Structures. Complex decisions must be decomposed into evaluable components, each of which can be assessed, critiqued, and revised independently. This allows decision makers to examine specific aspects of a problem, such as market viability or technical feasibility, without losing sight of the overall evaluation. It addresses the challenge of complex decision criteria (the most frequently cited difficulty in the survey) and the need for structured tracking of iterative decisions expressed by DM3.

DR4: Adjustable Autonomy (Human-in-the-Loop). Users must be able to control the level of AI involvement at each stage of the decision process. The system should support a range of modes, from a passive mode where the AI responds only when asked, to an active mode where it proactively identifies issues and proposes evaluations, to a deliberation mode where it facilitates a structured multi-perspective analysis. This reflects the consistent view across interviews that human oversight must be preserved, while acknowledging that the appropriate level of AI involvement may vary with context and user confidence.

DR5: Argumentation and Critique Capability. The system must be able to present arguments for and against decision options, challenge assumptions, and serve a devil’s advocate function when needed. This addresses the committee decision support use case identified by DM4, where subjective opinions need to be balanced with objective evidence. It also supports the reconciliation of conflicting stakeholder views and counteracts the HiPPO effect, where the most senior person’s view dominates regardless of evidence quality.

DR6: Integration of Experiential and Intuitive Knowledge. The system must provide ways to incorporate tacit knowledge that is not captured in formal data sources, including professional intuition, process experience, and contextual judgment. DM1 called for a collabora-

tive human-machine process that integrates experiential knowledge, and DM5 emphasized that interpersonal nuances and accumulated experience must be part of the decision process. Rather than treating human intuition as a bias to be corrected, the system should treat it as a valid and valuable input alongside data-driven analysis.

DR7: Privacy and Data Security by Design. Data protection must be built into the system architecture as a foundational principle, not added after the fact. This includes role-based access controls, secure handling of proprietary information, and clear data retention and usage policies. This requirement responds to the top concern in the survey (data protection, cited by 19 of 33 respondents) and the intellectual property concerns raised by DM14 and DM15.

5.2 The Incomplete Information Tension

The incomplete information tension introduced in Section 2.4 runs through all the design requirements and shapes how they should be read together. A practical response is the interaction between DR2 (gap identification) and DR6 (experiential knowledge integration): rather than concealing uncertainty, the system should make it explicit, state what is known, what is uncertain, and what is missing, and then actively invite human expertise to fill those gaps. In this way, incompleteness is reframed from a system limitation into a collaborative design feature, in which AI breadth and human judgment are treated as complementary rather than competing.

6. DESIGN PRINCIPLES

Table 6 summarizes how each design principle operationalizes one or more design requirements.

Table 6: Design principles and the requirements they operationalize

DP	Principle	Operationalizes
DP1	AI as analytical partner	DR4 (adjustable autonomy)
DP2	Transparency by default	DR1 (transparent reasoning chains)
DP3	Progressive disclosure	DR1, DR3 (reasoning chains; modular structure)
DP4	Multi-perspective deliberation	DR5, DR6 (argumentation and critique; experiential knowledge integration)
DP5	Graceful uncertainty handling	DR2, DR6 (gap identification; experiential knowledge as complement to uncertain data)
DP6	Incremental deployment	DR4, DR7 (adjustable autonomy; security by design)

While the design requirements (DR1–DR7) specify *what* the system must do, design principles articulate *how* the system should behave and what should guide its interaction with users. Six design principles are proposed, derived from the empirical findings and the literature on human-AI collaboration.

DP1: AI as Analytical Partner, Not Decision Maker. The system provides analysis, perspectives, and structured evaluations, but final decision authority always rests with the human. This principle reflects the consistent view across participants that AI should support but never replace human judgment, and operationalizes DR4 (adjustable autonomy) by drawing a clear boundary around the system’s role.

DP2: Transparency by Default. All reasoning processes, data sources, and confidence levels should be visible to the user without requiring special effort to access them. Rather than hiding complexity behind simple outputs, the system should present its analytical work in a way that invites scrutiny. This operationalizes DR1 (transparent reasoning chains) and responds directly to the consistent demand for traceability expressed across the interviews.

DP3: Progressive Disclosure of Complexity. Information should be accessible but presented in layers. The initial view offers a clear summary with key findings. Users who want more detail can drill down into underlying evidence, agent-level analyses, and raw data. Decision makers at different levels need different amounts of detail, and presenting everything at once can be as cognitively burdensome as providing no structure at all. This operationalizes DR1 and DR3 by making the system’s analytical work accessible without overwhelming the user.

DP4: Multi-Perspective Deliberation. The multi-agent architecture should be used to represent different functional or stakeholder viewpoints. Rather than producing a single monolithic recommendation, the system should make visible the perspectives it has considered, including areas of agreement and disagreement. This operationalizes DR5 (argumentation and critique) and DR6 (experiential knowledge integration) by treating both analytical outputs and practitioner expertise as inputs to deliberation, and directly addresses the cross-functional mediation need identified in the interviews.

DP5: Graceful Handling of Uncertainty. When the system encounters data gaps, conflicting information, or low-confidence situations, it should communicate this clearly rather than generating confident-sounding outputs. This means providing explicit confidence indicators, flagging areas of uncertainty, and distinguishing well-supported conclusions from speculative extrapolations. This principle directly addresses the incomplete information tension (Section 5.2) and operationalizes DR2 (gap identification) and DR6 (experiential knowledge integration), treating practitioner expertise as a complement to uncertain or incomplete data rather than a last resort.

DP6: Design for Incremental Deployment. The system architecture should support gradual introduction rather than all-or-nothing adoption. Individual analytical modules should be activatable independently, so that the system can deliver value with partial data integration before all organizational sources are connected. Autonomy levels (DR4) can be set conservatively at first and raised as trust develops. DM4 explicitly called for starting with a few clearly defined tasks and expanding scope gradually, an expectation with direct architectural implications. Incremental deployment also supports DR7 (privacy and data security): phased data integration allows security controls, access permissions, and data retention policies to be established and verified for each connected source before the next integration step proceeds (Saup et al., 2026).

7. DISCUSSION

7.1 Connecting to Existing Literature

The findings contribute to the literature on AI-supported decision making in innovation by showing that practitioners do not primarily ask for autonomous recommendations. They ask for systems that make reasoning visible, surface missing information, preserve human accountability, and help coordinate perspectives across functions. In this sense, the study provides contextual evidence for concerns already present in the responsible AI and human-AI collaboration literature, while specifying how those concerns manifest in innovation and product development settings. The results also suggest that some features commonly framed as technical enhancements, such as transparency or uncertainty handling, function in practice as conditions for adoption rather than optional extras.

The findings align with and extend two intersecting bodies of work: the design of AI-augmented decision-making (AIADM) systems, and the governance of LLM-based multi-agent architectures. The AIADM design principles of Herath Pathirannehelage et al. (2025) receive support in the innovation management context and are extended in two ways. First, when a multi-agent architecture is adopted, coordination requirements become more prominent: dis-

agreement between agents should be surfaced to users rather than resolved silently, and the coordination process itself should remain traceable. Second, innovation decisions rely heavily on tacit, experiential knowledge that is rarely captured in formal data, giving rise to DR6, a requirement not present in general AIADM design guidance. The “pilot-to-decision gap” documented by Saup et al. (2026) closely matches the pattern observed in the interviews; DR4 (adjustable autonomy) and DP6 (incremental deployment) together provide a trust-building pathway from exploratory use to embedded decision support.

The incomplete information tension connects to the automation bias literature (Romeo and Conti, 2025; Kücking et al., 2024): users tend to defer to AI outputs precisely when those outputs are least reliable. Guo et al. (2024b) formalizes this point by framing appropriate AI reliance as a function of AI reliability and task uncertainty, which suggests that DP5 (graceful uncertainty handling) should be treated as a prerequisite for effective human-AI complementarity rather than a secondary design concern.

7.2 Theoretical Contribution

This study offers three contributions. First, it provides empirical evidence about the conditions under which decision makers in innovation contexts consider AI-supported systems to be usable and trustworthy: reasoning traceability, explicit treatment of data gaps, adjustable levels of autonomy, and mechanisms for integrating cross-functional and experiential knowledge.

Second, these observations are translated into seven design requirements and six design principles. These should be understood as empirically derived design propositions rather than validated theory. Their value lies in making practitioner needs actionable for future system design, while acknowledging that evaluation through implemented artifacts is still needed.

Third, the study identifies and names the *incomplete information tension*: AI support appears most attractive when information is incomplete, yet those same conditions increase the risk that AI outputs are treated as more reliable than the evidence warrants. This concept is useful not because uncertainty is a new topic, but because it links transparency, uncertainty handling, and human oversight into a single design problem that future research can test more directly.

The study does not validate a specific multi-agent architecture. Rather, the findings suggest that a multi-agent configuration becomes most plausible where decision support must represent distinct functional perspectives, expose disagreement, or separate analytical tasks into traceable modules. A multi-agent system can address these needs more effectively than a single agent by isolating context windows, enabling specialized tool use, and generating conversational logs that naturally support transparency. The study’s contribution on this point is clarifying when such an architecture appears justified and what design obligations follow if it is adopted.

Figure 5 shows one possible reference architecture consistent with the requirements and principles derived in this study. The interaction and review interface operationalizes the AI-as-partner and progressive-disclosure principles (DP1, DP3). The orchestrator and autonomy manager handle modular task decomposition and adjustable autonomy (DR3, DR4, DP6). The analytical role layer represents the distinct functional perspectives that practitioners identified as necessary, and the synthesis layer makes disagreement, evidence integration, and gap identification explicit (DR2, DR5, DP4, DP5). Human experiential input is included as an explicit input, not a fallback (DR6). Reasoning traceability, provenance, privacy, and auditability (DR1, DR7, DP2) are shown as cross-cutting controls.

To illustrate the architecture concretely, consider a make-or-buy decision for a new software component. The user submits the request to the orchestrator, which delegates it to specialized agents. A financial agent evaluates internal development costs; a technical agent assesses external vendor capabilities. If the financial agent finds a cost advantage but the technical agent

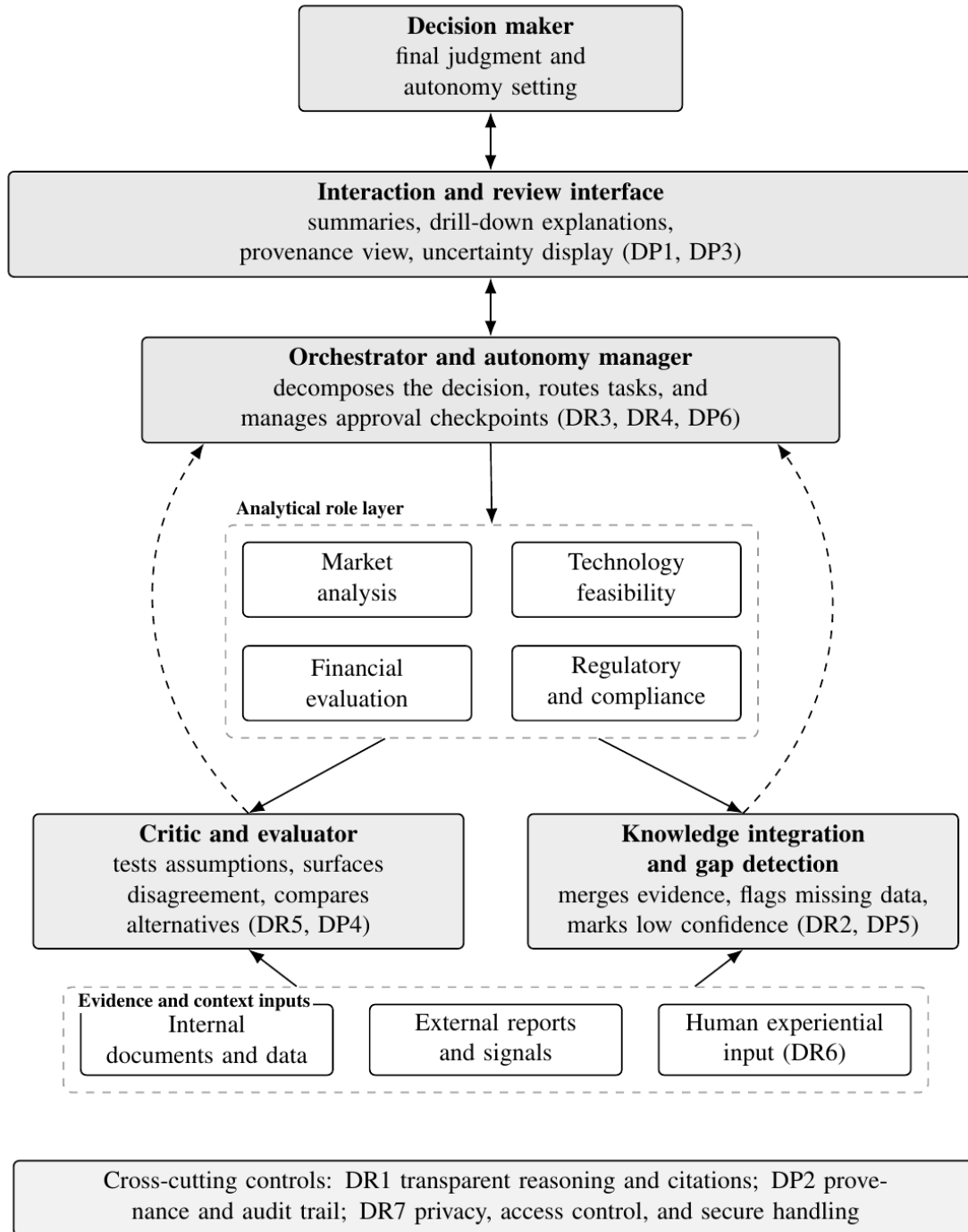


Figure 5: Illustrative LLM-based multi-agent reference architecture consistent with the proposed design requirements and principles (DR1–DR7, DP1–DP6). The diagram is conceptual and analytically derived; it has not been implemented or empirically validated.

flags a data quality gap about the vendor, the synthesis layer surfaces this conflict rather than forcing a single recommendation, and prompts the user to assess the acceptable level of risk. The figure is intended as a synthesis device that makes the study’s multi-agent implications concrete without claiming artifact-level validation.

7.3 Managerial Implications

For innovation managers, two practical points stand out. First, feasibility assessment and make-or-buy analysis are among the most suitable initial use cases: decision criteria are relatively well-defined, relevant information sources are identifiable, and the output is a comparative evaluation rather than an open-ended recommendation. Starting with structured problems reduces the risk of early adoption failures that can erode organizational confidence. Second, polished AI outputs can make experienced users less critical, not more (Kücking et al., 2024). Organizations should therefore invest in what may be called *critical collaboration skills* (cf. Steyvers and Kumar, 2024): the ability to interrogate AI reasoning, identify evidence gaps, and exercise independent judgment when stated confidence does not match data quality. DR1 and DP2 are designed to make such scrutiny possible; building the organizational disposition to actually use that capacity is a managerial responsibility.

For technology developers, transparency and traceability (DR1, DP2) should be treated as architectural foundations, not afterthoughts. In regulated industries such as medical devices and pharmaceuticals, every AI-assisted interaction may be subject to audit, and systems that cannot show their reasoning will not be adopted precisely where decision support is most needed. Data gap identification (DR2) was the most consistently requested capability in the interviews and among the least present in current LLM applications, making it a core engineering priority. Adjustable autonomy (DR4) is equally important: a fixed level of AI involvement will be too high for cautious users and too low for confident ones. At the governance level, data protection safeguards (DR7) must be built in before deployment, not added later; the system’s formal role in the decision process should also be defined before rollout to avoid accountability ambiguity when AI is involved in consequential decisions.

7.4 Limitations and Future Research

Three scope limitations should be considered when interpreting the findings. The geographic concentration in German organizations means that cultural attitudes toward algorithmic authority and local regulatory norms may have shaped the results; cross-cultural replication is needed to identify which requirements are broadly generalizable. Both samples are modest in size and purposively selected, making the findings rich in depth but not statistically representative; larger surveys would help identify which patterns hold at scale. Participants were likely more positively oriented toward AI than the broader population of innovation managers due to self-selection in consortium-based recruitment. The study also focuses primarily on physical product development and manufacturing-adjacent innovation; whether the design requirements hold equally in service innovation or software-as-a-service contexts remains to be examined.

The most consequential limitation is that the study derives design guidance without evaluating an implemented system. The proposed requirements and principles are analytically grounded but empirically untested at the artifact level. The framework also does not fully account for current technical constraints in multi-agent architectures, such as computational latency, hallucination risks, and token processing costs. Future work should therefore move from concept derivation to prototype development and situated evaluation, testing not only technical feasibility but whether the proposed features improve decision quality and perceived human control in real settings. Longitudinal designs are particularly important, as human-AI collab-

oration quality is unlikely to be stable at first exposure and may change substantially as users develop familiarity with both the system’s capabilities and its failure modes.

8. CONCLUSION

Drawing on a survey of 33 practitioners and 16 semi-structured interviews across German firms, this study identified three structural breakdowns in innovation decision processes: information fragmentation, absent reasoning structure, and the need to act under incomplete evidence. In response, an empirically grounded design framework is proposed that positions AI as an analytical partner, makes uncertainty explicit, and keeps human judgment central. A key insight is the *incomplete information tension*: AI support is most needed precisely where data are scarce, yet those same conditions heighten the risk of overreliance on outputs whose underlying uncertainty remains hidden. While the proposed framework is analytically grounded, future research must develop and evaluate functional prototypes in real organizational settings to validate whether these design features improve decision quality and preserve meaningful human control in practice.

REFERENCES

- Bastone, A., Mandiello, A., and Zeuli, F. (2026). Generative artificial intelligence and decision-making: Evidence from a participant observation with latent entrepreneurs. *European Journal of Innovation Management*, pages 1–21.
- Bouschery, S. G., Blazeovic, V., and Piller, F. T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management*, 40(2):139–153.
- Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.
- Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismanni, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., and Maharaj, T. (2023). Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’23)*, pages 651–666.
- Chiarello, F., Giordano, V., Spada, I., Barandoni, S., and Fantoni, G. (2024). Future applications of generative large language models: A data-driven case study on ChatGPT. *Technovation*, 133:103002.
- Creswell, J. W. and Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications, Thousand Oaks, CA, 5th edition.
- Füller, J., Tekic, Z., and Hutter, K. (2024). Rethinking innovation management: How ai is changing the way we innovate. *Journal of Applied Behavioral Science*, 60(4):603–612.
- Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., and Zhang, X. (2024a). Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680.
- Guo, Z., Wu, Y., Hartline, J. D., and Hullman, J. (2024b). A decision theoretic framework for measuring AI reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’24)*, pages 1310–1321.

- Herath Pathirannehelage, S., Shrestha, Y. R., and von Krogh, G. (2025). Design principles for artificial intelligence-augmented decision making: An action design research study. *European Journal of Information Systems*, 34(2):207–229.
- Hettrich, B., Krings, N., and Kock, A. (2025). Bridging the expertise gap: The role of generative ai in supporting project planning tasks for novices and professionals. *Creativity and Innovation Management*.
- Hosseini, S. and Seilani, H. (2025). The role of agentic ai in shaping a smart future: A systematic review. *Array*.
- Johnson, R. B. and Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational Researcher*, 33(7):14–26.
- Krueger, E., Carpenter, T., Leach, K., and Weimer, J. (2025). Towards an llm-based system for generating and validating product requirements. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering (FSE Companion '25)*, pages 1538–1539.
- Kücking, F., Hübner, U., Przysucha, M., Hannemann, N., Kutza, J.-O., Moelleken, M., Erfurt-Berge, C., Dissemond, J., Babitsch, B., and Busch, D. (2024). Automation bias in ai-decision support: Results from an empirical study. *Studies in Health Technology and Informatics*, 317:298–304.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Mariani, M. M. and Dwivedi, Y. K. (2024). Generative artificial intelligence in innovation management: A preview of future research developments. *Journal of Business Research*, 175:114542.
- Osterwalder, A., Pigneur, Y., Bernarda, G., and Smith, A. (2014). *Value Proposition Design*. John Wiley & Sons, Hoboken, NJ.
- Patton, M. Q. (2015). *Qualitative Research & Evaluation Methods*. SAGE Publications, Thousand Oaks, CA, 4th edition.
- Peng, L., Li, D., Zhang, Z., Zhang, T., Huang, A., Yang, S., and Hu, Y. (2024). Human-ai collaboration: Unraveling the effects of user proficiency and ai agent capability in intelligent decision support systems. *International Journal of Industrial Ergonomics*, 103:103629.
- Romeo, G. and Conti, D. (2025). Exploring automation bias in human-ai collaboration: A review and implications for explainable ai. *AI & Society*.
- Sami, M. A., Zhang, Z., Waseem, M., Kemell, K.-K., Rasheed, Z., Herda, T., and Abrahamsson, P. (2026). Bridging humans and llms: Investigating human-ai collaboration in multi-agent requirements analysis for organizational ai adoption. *e-Informatica Software Engineering Journal*, 20(1).
- Saup, T.-O., Asghar, J., Kanbach, D. K., and Kraus, S. (2026). From pilots to decision systems: Embedding generative AI into strategic decision-making through a socio-technical and governance lens. *Journal of Decision Systems*, 35(1):2597835.
- Schuh, G., Lenders, M., and Hieber, S. (2011). Lean innovation: Introducing value systems

- to product development. *International Journal of Innovation and Technology Management*, 8(1):41–54.
- Schwarz, G., Christensen, T., and Zhu, K. (2022). Bounded rationality, satisficing, artificial intelligence, and decision-making in public organizations: The contributions of herbert simon. *Public Administration Review*, 82(5):902–904.
- Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, 69(1):99–118.
- Steyvers, M. and Kumar, A. (2024). Three challenges for ai-assisted decision-making. *Perspectives on Psychological Science*, 19(5):722–734.
- Sun, C., Huang, S., and Pompili, D. (2025). Llm-based multi-agent decision-making: Challenges and future directions. *IEEE Robotics and Automation Letters*, 10(6):5681–5688.
- Varnes, C. (2025). Unpacking the context of decision criteria: Four types of enacted and sustained competitiveness. In *Proceedings of the 31st Innovation and Product Development Management Conference (IPDMC 2024)*.
- Witkowski, A. and Wodecki, A. (2025). Where does ai play a major role in the new product development and product management process? *Management Review Quarterly*.
- Yang, A. and Yang, T. A. (2024). Social dangers of generative artificial intelligence: Review and guidelines. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, pages 654–658.

A. SURVEY INSTRUMENT SUMMARY

The online survey was administered as part of the *InnoHub Generative AI* programme. Table 7 provides an overview of the key survey items and the constructs they were designed to measure.

Table 7: Survey instrument: key constructs, items, and response formats

#	Construct	Item / Focus	Format
1	AI interest	Interest in AI for leadership and management	Likert (5-pt)
2	AI knowledge	Self-assessed AI knowledge in a management context	Likert (5-pt)
3	AI familiarity	Familiarity with AI-based and multi-agent DSS	Likert (5-pt)
4	Decision frequency	Frequency of strategic decision making	Single choice
5	Decision factors	Key factors influencing strategic decisions	Single choice
6	Methods used	Traditional methods used (SWOT, CBR, data analysis, 5-Why, MCDM)	Multi-select
7	Challenges	Challenges in SDM (complex criteria, time pressure, data gaps, tool limitations, uncertainty)	Multi-select
8	Concerns	AI adoption concerns (transparency, trust, data protection, integration)	Multi-select
9	Preferred insights	Desired AI insight types ranked (trend, risk, feasibility, customer, competitor)	Ranking (1–5)
10	Current tools	Tools used for decision support; perceived limitations	Open-ended
11	Future expectations	Expected role of AI in decision making (5-year horizon)	Open-ended

B. INTERVIEW GUIDE SUMMARY

The semi-structured interview protocol was organised around four main question areas, with probing questions used to elicit deeper insights as needed. Table 8 summarises the protocol.

Table 8: Interview guide: main question areas and sample probes

Question Area	Sample Questions and Probes
1. Role and decision characteristics	Describe your role and the types of strategic decisions you are involved in. How are these decisions typically made? Who is involved?
2. Challenges and tool use	What are the main challenges in these decision processes? Which tools are currently used, and what are their limitations? Where do breakdowns or inefficiencies tend to occur?
3. Use cases for LLM-MAS	Where do you see the greatest potential for AI-based multi-agent systems in your decision processes? What tasks would you want such a system to handle?
4. Requirements and concerns	What requirements would an AI-based decision support system need to meet for you to trust and use it? What concerns do you have about such systems?